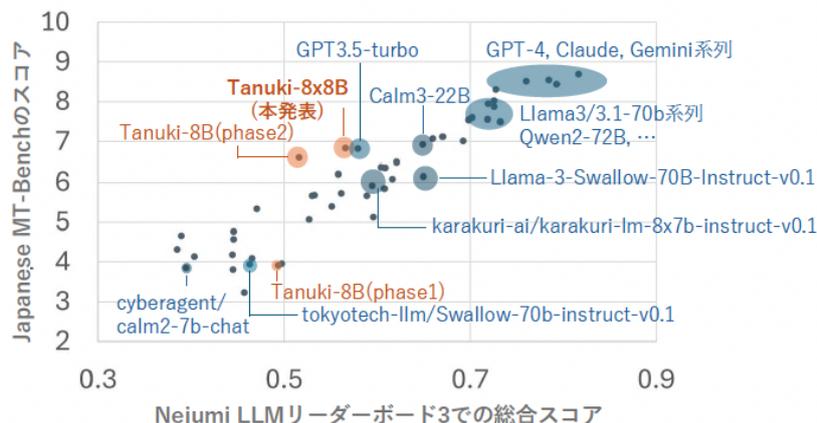


東京大学松尾・岩澤研究室 GENIAC プロジェクトにおいて、大規模言語モデル 「Tanuki-8×8B」を開発・公開

発表のポイント

- ◆ 経産省及びNEDOが進める日本国内の生成 AI 基盤モデル開発を推進する「GENIAC」プロジェクトにおいて、松尾・岩澤研究室が「Tanuki-8×8B」を開発・公開。
- ◆ 本モデルは、フルスクラッチで開発されており、対話、作文能力を評価する指標「Japanese MT-Bench」において「GPT-3.5 Turbo」と同等以上の性能を達成。
- ◆ Apache License 2.0 のライセンスに基づき、研究および商業目的での自由な利用が可能。「Tanuki-8×8B」の軽量版である、「Tanuki-8B」をチャット形式で利用できるデモも公開。



本モデルの Nejumi LLM リーダーボード 3 における評価

発表内容

東京大学大学院工学系研究科技術経営戦略学専攻 松尾・岩澤研究室（以下「松尾研」）は、経済産業省及び国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）が推進する、国内の生成 AI の開発力を強化するためのプロジェクト「GENIAC (Generative AI Accelerator Challenge)」(注1)において、大規模言語モデル「Tanuki-8×8B」を開発し、公開いたしました。

〈開発の背景〉

本活動は、日本国内の生成 AI 基盤モデル開発を推進する「GENIAC」プロジェクトにおいて、2023年8月に公開した100億パラメータサイズのLLM「[Web1ab-10B](#)」の開発経験をベースに、LLM開発を進めるものです。

松尾研が提供する[大規模言語モデル講座](#)（2023年8月開催、2000名以上が受講）の修了生及び一般公募によって集まった有志のメンバー（民間企業・研究者・学生で構成）が、それぞれの知見を取り入れながら開発に取り組みました。

開発工程は2つのフェーズに分かれており、Phase1（注2）では7チームに分かれてコンペティション形式で開発を進め、Phase2ではPhase1の優勝チームが更に大規模なモデル開発に挑戦しました。「Tanuki-8×8B」は本取り組みの結果開発され、公開するものです。

松尾研 GENIAC プロジェクトについては下記をご覧ください。

https://weblab.t.u-tokyo.ac.jp/geniac_llm/

「Tanuki-8×8B」の概要

「Tanuki-8×8B」はPhase1で構築された8Bモデルを8つに複製し、それぞれを専門家モデルとして分化・連携させることで動作するように効率的に追加学習されたモデル（注3）です。本モデルは、フルスクラッチで開発されており、作文、会話を評価する指標「Japanese MT-Bench」においては「GPT-3.5 Turbo」と同等以上の性能を達成しています。モデル名は「日本らしく、親しみを覚える動物の名前」というテーマでチームメンバーで案を出し、投票で決定しました。

■ 開発モデル公開 URL

本モデルは Apache License 2.0 のライセンスに基づき、研究および商業目的での自由な利用が可能です。

Tanuki-8x8B : <https://huggingface.co/weblab-GENIAC/Tanuki-8x8B-dpo-v1.0>

Tanuki-8B : <https://huggingface.co/weblab-GENIAC/Tanuki-8B-dpo-v1.0>

■ デモ公開 URL

「Tanuki-8×8B」の軽量版である、「Tanuki-8B」をチャット形式で利用できるデモを下記 URL で公開しております。下記 URL にアクセスし実際の会話をお試しく下さい。

<https://huggingface.co/spaces/weblab-GENIAC/Tanuki-8B-dpo-v1.0>

※期間限定公開・終了日時未定

Tanuki-8B-dpo-v1.0
Duplicate Space for private use

Chatbot

日本で有名なものと言えば

日本で有名なものは多岐にわたりますが、いくつか代表的なものを紹介します。まず、アニメとマンガは世界中で愛されており、スタジオジブリの作品や「ワンピース」、「ナルト」、「鬼滅の刃」などが特に有名です。また、アニメ映画としても「新世紀エヴァンゲリオン」や「君の名は。」が国内外で高い評価を受けています。

次に、漫画も日本独自の文化として非常に有名です。「ドラゴンボール」や「ワンピース」、「スラムダンク」など、多くの名作が存在します。さらに、マンガだけでなく、アニメだけでなく実写映画やドラマも人気で、「半沢直樹」や「逃げ恥」といったドラマが社会現象を巻き起こしました。

食べ物に関しても、日本は世界中で知られています。寿司やラーメン、天ぷら、お好み焼きなど、寿司は特に高級料理として世界中で愛されています。また、和菓子や抹茶も日本を代表する食文化の一部です。

観光名所も数多く、京都の金閣寺や清水寺、東京の浅草寺、大阪の大阪城など、歴史と現代が融合した魅力的なスポットが豊富です。富士山も日本の象徴的な自然景観として広く知られています。

さらに、テクノロジー分野でも日本は世界をリードしています。ソニーやパナソニック、トヨタなどの企業はグローバルに活動しており、特に自動車技術やエレクトロニクス製品で高い評価を受けています。

最後に、伝統文化も忘れてはなりません。茶道、花道（生け花）、能楽、歌舞伎など、日本独自の芸術や儀式が今も大切に受け継がれています。これらの影響は日常生活や観光にも深く根付いており、日本を訪れる多くの観光客がその美しさと深みを堪能しています。

以上のように、日本には多様な分野で有名なものが数多く存在し、それぞれが独自の魅力を持っています。

Retry Undo Clear

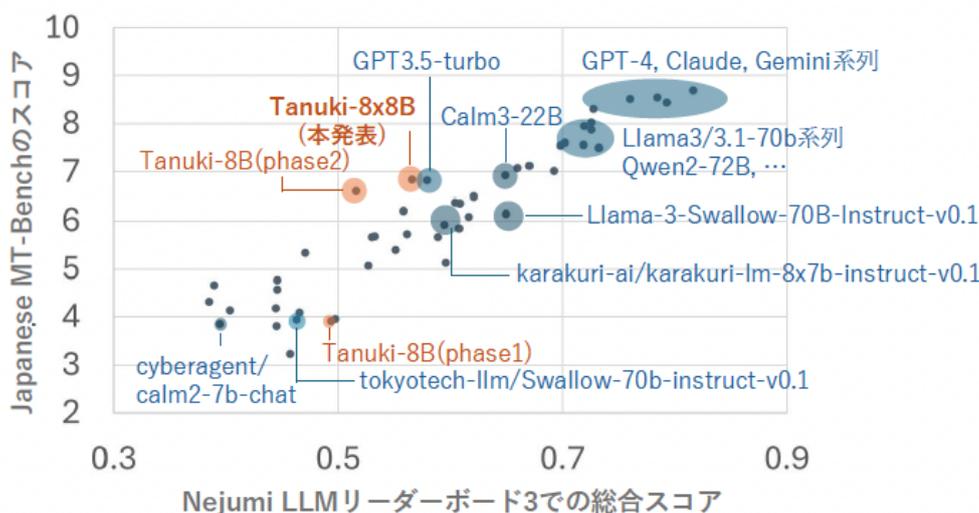
デモ画面

「Tanuki-8×8B」の特徴

「Tanuki-8×8B」は文章の作文や対話を中心に学習しており、当該能力を測る「Japanese MT-Bench」では事前学習からフルスクラッチで開発を行ったモデルとして、GPT-3.5 Turbo と同等の性能を達成しています。

■ Nejumi LLM リーダーボード 3 における評価

はじめに、大規模言語モデルの総合的な日本語能力を評価するための最新（注4）のベンチマークシステムである [Nejumi LLM リーダーボード 3](#) のベンチマークプログラムを用いてモデル性能を評価し、既存のモデル群（注5）と比較しました。



橙色が本プロジェクトで開発されたモデル、青色が他モデルを表す

Nejumi LLM リーダーボード 3 は言語モデルの文章作成、対話能力のほか、知識や安全性など、さまざまな観点から性能を評価するシステムです。本ベンチマークにおける「Tanuki-8×8B」の総合スコアは 0.57/1.00 で、OpenAI の GPT-3.5 Turbo (0.58/1.00) と同程度の性能であることが分かりました。

今回の開発で最も力を入れた点は、モデルの作文・対話能力の向上です。この能力を客観的に評価する Japanese MT-Bench では、10 点満点中 7 点程度のスコアが得られました。この成績は、GPT-3.5 Turbo (7 点) や CyberAgentLM3-22B-Chat (7 点) と同水準（注6）です。

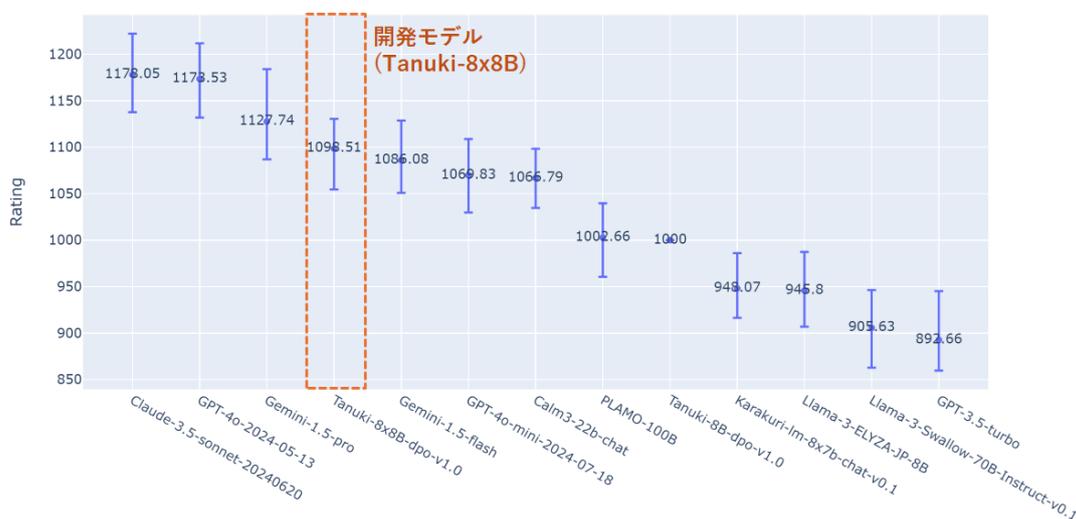
■ 人手での評価

Nejumi LLM リーダーボード 3 での計測に加えて、人手でのモデル評価を行いました。モデルの性能評価に用いた [Nejumi LLM リーダーボード 3](#) は、さまざまなトピックに関するモデルの出力を、アルゴリズムや GPT-4 を用いた採点によって全自動で評価する枠組みです。客観性や再現性に優れた計測手法ではありますが、評価には人間のフィードバックや実際の対話評価が含まれないため、ベンチマークスコアとユーザーの使用感には乖離が生じる可能性があります。

そこで本プロジェクトでは、より実践的な観点からモデルの作文・対話能力を評価することを目的として、ユーザーの質問に対する種々のモデルの回答の優劣をブラインドテストの形で

競わせる、いわゆる [ChatBot Arena](#) として広く知られるシステムと同じ原理で追加評価を実施しました。具体的には、ユーザーの質問に対し、ランダムに選ばれた 2 つの言語モデルの出力の優劣を、人間が評価する試験を実施しました（評価期間：2024 年 8 月 19～24 日）。開発チームに加え、松尾・岩澤研の LLM コミュニティに参加するメンバーが評価に参加し、2000 件以上の対話データを収集しました（後日公開予定）。

国内で開発、あるいは追加学習された代表的な高性能モデルに加え、GPT-3.5/4、Gemini-1.5、Claude-3.5 などの商用 LLM を含めた 13 種類のモデルの出力を比較評価しました。集計の結果、「Tanuki-8×8B」の性能は GPT4o、Gemini-1.5-pro、Claude-3.5-sonnet に次ぐ順位に相当することが分かりました。GPT-4o-mini、Gemini-1.5-flash、CyberAgentLM3-22B-Chat と同レベルの性能帯に位置しています。一連の評価より、ユーザーとの対話を想定した実践的なフェーズにおいて、開発した 8x8B モデルが優れた出力性能を示すことが分かりました。



ChatBot Arena の Bradley-Terry model によってモデルの優劣をレーティングした結果

(Tanuki-8B の性能を基準値 (1000) に設定)

モデル名	モデル重みの公開	レーティング	勝率	対戦回数	Japanese MT-Bench平均点	Nejumi leader board 3総合点
GPT-4o-2024-05-13	非公開	1178	0.57	297	8.6	0.78
Claude-3.5-sonnet-20240620	非公開	1173	0.56	272	8.7	0.82
Gemini-1.5-pro	非公開	1128	0.50	208	7.9	0.73
Tanuki-8x8B-dpo-v1.0	オープン	1099	0.41	382	7.0	0.57
GPT-4o-mini-2024-07-18	非公開	1086	0.41	303	8.3	0.72
Gemini-1.5-flash	非公開	1070	0.41	318	7.6	0.70
Calm3-22b-chat	オープン	1067	0.36	392	6.9	0.65
PLAMO-100B	非公開	1002	0.29	297	(報告無し)	(報告無し)
Tanuki-8B-dpo-v1.0	オープン	1000	0.28	379	6.6	0.52
Karakuri-lm-8x7b-chat-v0.1	オープン	948	0.21	329	5.8	0.60
Llama-3-ELYZA-JP-8B	オープン	945	0.17	282	6.1	0.62
Llama-3-Swallow-70B-Instruct-v0.1	オープン	906	0.16	348	6.2	0.65
GPT-3.5-turbo	非公開	893	0.17	291	6.8	0.58

モデルの諸性能の比較

本モデルに残された課題として、Japanese MT-Bench の結果も含めた総合的な推論性能は GPT-4o、Gemini-1.5-pro、Claude-3.5-sonnet などの海外の最先端モデルにはまだ追いついていない点などが挙げられます。一方で「Tanuki-8×8B」は、海外モデルとは異なる長所を持つことも明らかになりつつあります。海外モデルはどちらかというと無機質で形式的な返答をする傾向にありますが、それとは対照的に、当該モデルは共感性や思いやりのある返答や、自然な言葉遣いでの作文が得意でした。

Gemini-1.5-proの回答

Tanuki8×8Bの回答

<p>掃除がめんどくさいのですが、どんなモチベーションでやればよいのでしょうか。</p>	<p>掃除がめんどくさいと感じるのはよくあることですが、いくつかのモチベーションを見つけることで、そのプロセスを少しでも楽しく、または効率的にすることができます。以下にいくつかの方法を提案します。</p> <ol style="list-style-type: none"> 目標を設定する: <ul style="list-style-type: none"> - 短期的な目標と長期的な目標を設定しましょう。例えば、「今日はリビングのテーブルを拭く」や「毎週末にキッチン掃除をする」といった具体的な目標を立てると、達成感が得られやすくなります。 環境を整える: <ul style="list-style-type: none"> - 掃除しやすい環境を作ることも重要です。整理整頓された空間は、掃除をする際の負担を軽減します。不要な物を捨てたり、収納スペースを見直したりすることで、掃除が楽になります。 音楽やポッドキャストを活用する: <ul style="list-style-type: none"> - 好きな音楽を流しながら掃除をすると、気分が上がる場合があります。また、掃除に関するポッドキャストやオーディオブックを聞くことで、掃除の時間が学びの時間にもなります。 報酬システムを導入する: <ul style="list-style-type: none"> - 掃除が終わった後に自分へのご褒美を設定するのも効果的です。例えば、好きなスイーツを食べる、映画を観る、趣味の時間を持つなど、小さな報酬を用意するとモチベーションが高まります。 パートナーや家族と一緒にやる: <ul style="list-style-type: none"> - 誰かと一緒に掃除をすることで、楽しく作業を進めることができます。お互いに励まし合いながら進めると、孤独感が軽減されます。 	<p>掃除、面倒ですよ...わかります！でも、終わったら気持ちいいのも事実。そこで、面倒くさいを吹き飛ばすためのモチベーションアップのコツをいくつか紹介します！</p> <p>**1. ご褒美作戦**</p> <p>***小さな目標達成ごとに自分にご褒美をあげましょう！** 例えば、「今日はキッチンを磨いたら、美味しいケーキを食べる」「クローゼットの整理が終わったら、気になっていた映画を観る」など。 *ゲーム感覚で楽しめば、掃除の時間が待ち遠しくなるかも？</p> <p>**2. 快適空間をイメージ**</p> <p>***掃除後のすっきりとした部屋を想像してみましょう！** 「ピカピカの床でヨガができたら気持ちいいな」「整理整頓された机で集中して作業したいな」など。 *イメージを膨らませることで、行動に移す力が湧いてきます。</p> <p>**3. 音楽やアロマの力を借りる**</p> <p>***アップテンポな音楽をかけながら、掃除をすれば、楽しくはかどります！** 好きなアーティストのライブ映像などを流せば、掃除もエクササイズ感覚で楽しめるかも？ ***リラックス効果のあるアロマを焚けば、掃除しながら癒やしの時間を過ごせます。**</p> <p>**4. 「ながら掃除」を取り入れる**</p>
--	--	--

Tanuki8×8Bの回答

Claude-3.5-sonnetの回答

<p>テストで95点でもミスがあったと言って親は褒めてくれません。</p>	<p>お子さんの気持ちはよくわかります。95点という素晴らしい成績を取ったのに、親から褒められないのは辛いですよね。以下のようなアドバイスが役立つかもしれません：</p> <ol style="list-style-type: none"> 自分の努力と成果を認めましょう。95点は素晴らしい成績です。 親とオープンに話し合ってみてください。あなたの気持ちを伝えることが大切です。 親の期待が高すぎると感じるなら、それについても話し合ってみましょう。 完璧を求めすぎず、努力の過程を大切にすることの重要性を理解してもらいましょう。 自分の成長に焦点を当て、他人の評価にとらわれすぎないようにしましょう。 信頼できる先生や友人に相談するのも良いかもしれません。 自分を褒める習慣をつけましょう。自己肯定感を高めることが大切です。 <p>親の反応に落胆するのは自然なことですが、あなたの努力と成果は価値があります。自信を持ち続けてください。</p>	<p>テストで95点を取ったことは素晴らしい成果ですが、親がその点数を褒めてくれない理由にはいくつかの可能性があります。以下にステップバイステップで説明します。</p> <ol style="list-style-type: none"> 期待値の確認: <ul style="list-style-type: none"> - まず、親がどのような期待を持っていたのかを確認することが重要です。例えば、親が「常に満点を目指してほしい」と思っている場合、95点では満足できないかもしれません。 ミスの詳細確認: <ul style="list-style-type: none"> - ミスが具体的に何だったのかを理解しましょう。ミスの内容によっては、親がそのミスを重要視している可能性があります。例えば、計算ミスやスペルミス、論理的な誤りなど、どの部分が問題だったのかを明確にします。 コミュニケーション: <ul style="list-style-type: none"> - 親とオープンなコミュニケーションを取ることが大切です。テストの結果について話し合い、なぜ親が満足していないかを尋ねてみましょう。 改善点の共有: <ul style="list-style-type: none"> - ミスを改善するための具体的な方法を提案します。例えば、「次回は計算ミスを減らすために、見直しの時間を増やそう」とか「論理的な誤りを防ぐために、問題を解く前に全体の流れを確認しよう」といった具体的なアドバイスを提供します。 ポジティブなフィードバック: <ul style="list-style-type: none"> - 親が満足していない理由が理解できたら、その上でポジティブなフィードバックを求めます。「今回は95点とれたことを誇りに思うよ。でも、次回はもっと良い点を取れる
---------------------------------------	---	--

GPT-4oの回答

Tanuki8×8Bの回答

人生最後の日に何をしたいですか？	GPT-4oの回答	Tanuki8×8Bの回答
	<p>人生最後の日に何をしたいかという問いは、非常に個人的で深い考えを呼び起こします。答えは人それぞれ異なるでしょうが、多くの人が共感するであろういくつかの一般的な願望を以下に挙げてみます。</p> <ol style="list-style-type: none"> “大切な人と過ごす” 家族や友人、愛する人たちと最後の時間を共有すること、一緒に話したり、過去の思い出を振り返ったり、感謝の気持ちを伝えることに時間を割きましょう。 “好きな場所を訪れる” 特別な思い出が詰まった場所や、今まで行きかけた場所を訪れること。自然の中や海辺、山など、心が落ち着く場所で最後の一日を過ごすことも考えられます。 “最後の言葉を残す” 家族や友人に手紙を書いたり、ビデオメッセージを残したりすることで、今後も彼らに対する思いや言葉を伝えることができます。 “お気に入りのことをする” 好きな食べ物を食べたり、お気に入りの映画や音楽を楽しんだりすることで、自己満足感を得る方法もあります。 “感謝と許し” これまでの人生でお世話になった人々に感謝の言葉を伝え、未解決の問題やわだかまりがあれば、許しを求めたり、自ら許したりすることも重要です。 “リラックスと瞑想” 最後の一日を静かで穏やかな気持ちで過ごすために、瞑想や深呼吸を取り入れることも一案です。 “スピリチュアルな活動” 宗教や信仰を持っている人にとっては、祈りや儀式を通じて心の平安を求めることも大切です。 <p>結局のところ、最後の日に何をしたいかは、その人の人生観や価値観によります。ご自身が最も大切だと思うこと、最も時間を割きたい人々や活動に焦点を当てるのが、最善の選択となるでしょう。</p>	<p>人生最後の日に何をしたいかを考えることは、非常に個人的で深いテーマです。この問いに対する答えは、個々の価値観や経験、信念によって大きく異なるでしょう。以下はいくつかの一般的な視点と、それに基づく回答の例です。</p> <ol style="list-style-type: none"> 家族や友人と過ごす - 多くの人にとって、大切な人たちと一緒に過ごすことが最も価値のある瞬間かもしれません。愛する人々との思い出を共有し、感謝の気持ちを伝えることができるでしょう。 自然の中でリラックスする - 美しい風景や静かな場所で最後の時間を過ごすことも考えられます。自然の中で心を落ち着け、自分自身と向き合う時間を持つことで、満足感を得られるかもしれません。 特別な活動をする - これまで挑戦できなかったことや、やり残したことを実行することも一つの方法です。例えば、旅行、趣味の活動、ボランティア活動など、自分が情熱を持っていることに時間を費やすことができます。 感謝の気持ちを表す - 人生を振り返り、感謝の気持ちを伝えることも重要です。家族や友人、同僚、そして自分自身に対して、これまで支えてくれたすべての人々に感謝の言葉を伝えることで、心の平安を得ることができるでしょう。 瞑想や祈り - 内省的な時間を持ち、自分自身と向き合い、精神的な平和を追求することも一つの選択です。瞑想や祈りを通じて、人生の意味や目的について深く考えることができるかもしれません。 <p>最終的には、自分自身が何を大切に、どのような価値観を持っているかによって、最適な過ごし方が決まるでしょう。人生最後の日は、その人が本当に大切にしているものを見つけ、それに基づいて行動することが最も意味のある瞬間となるでしょう。</p>

今回のプロジェクトを通して集積したオープンな知見をもとにモデル開発を継続して行うことで、日本ならではのオリジナリティや競争力を兼ね備えた LLM 群が誕生することが期待されます。

○関連情報：

本モデルの開発過程やナレッジは、下記プロジェクトページや Zenn ブログ記事にてオープンに公開されております。詳細は下記をご覧ください。

GENIAC 松尾研 LLM 開発プロジェクト特設ページ：

https://weblab.t.u-tokyo.ac.jp/geniac_llm/

Zenn ブログ記事：

<https://zenn.dev/p/matsuolab>

注釈

(注 1) 国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) 「ポスト 5G 情報通信システム基盤強化研究開発事業／ポスト 5G 情報通信システムの開発」事業。経済産業省が主導する基盤モデルの開発に必要な計算資源に関する支援や関係者間の連携を促す「GENIAC」プロジェクトの一環として採択事業者に一定の計算資源に関わる助成を行うもの。

GENIAC の詳細はこちら：

<https://www.meti.go.jp/press/2023/02/20240202003/20240202003.html>

https://www.meti.go.jp/policy/mono_info_service/geniac/index.html

(注 2) Phase1 で各チームが開発したモデル、コードも併せて下記にて公開。

モデル (HuggingFace) : <https://huggingface.co/weblab-GENIAC>

コード (GitHub) : https://github.com/matsuolab/nedo_project_code

(注 3) Mixture of Experts (MoE) と呼ばれるアーキテクチャを、アップサイクリングと呼ばれる手法で実現。総パラメータは数 47B、アクティブパラメータ数は 13B。アップサイクリングについての参考論文は下記。

<https://arxiv.org/abs/2406.06563>

<https://arxiv.org/abs/2212.05055>

(注 4) 2024 年 8 月時点での情報。

(注 5) 開発モデルの性能はプロジェクト内部で評価。既存モデルの性能については、当該サイトから 2024 年 8 月 13 日にダウンロードしたデータを使用。

(注 6) ベンチマークプログラムの仕様上、同一のモデルであっても、測定のたびに 0.1-0.2 点程度のスコア変動が生じるため、モデル間の順位が変動することを確認しています。

問合せ先

東京大学大学院工学系研究科 松尾・岩澤研究室 広報担当